SOCIAL NETWORK ANALYSIS REPORT “AMBER HEARD”

Machine Learning Approach Over 2020-2021

JUNE 30, 2021

MOHAMMED A. YASSIN
[redacted]

First Data Processing and Machine training

Machine Learning model

- The Machine learning model used was made by another Data scientist at github with accuracy of about 85%

- The training data used by the owner of model was from Kaggle twitter-bot-dataset unfortunately the dataset was unavailable now so I downloaded new dataset from Botometer website

Features used by model :

1. Verified -> account verification 0/1 value
2. hour_created -> hour the tweet was made datetime value
3. geo_enabled -> if geo location was enabled 0/1 value
4. default_profile -> When true, indicates that the user has not altered the theme or background of their user profile 0/1 value
5. Default_profile_image ->When true, indicates that the user has not uploaded their own profile image and a default image is used instead 0/1 value
6. Favourites_count -> number of tweets user liked throughout lifetime int value
7. Followers_count -> number of followers int value
8. Friends count -> number of following accounts int value
9. Statuses count-> number of tweets/retweets made by accounts int value
10. Average tweet per day -> statuses count/day_age of account float value
11. Network
12. Tweet to followers
13. follower acq rate
14. Friends acq rate

Data collection and Wrangling

Training Data :-

the data I got in order to train Machine Learning model (ML) was from Botometer dataset “cresci-stock-2018” It was a Json file and a txt file of id and if it’s bot or not I did some wrangling like changing name of some columns forming new columns from other original columns eg (one column called “account_age_days” was by taking the difference created_at and created_at_1 t(time of recent tweet))

Note that the data is 0 – 1 type I.e. it’s either a bot or not bot without giving the probability of being bot in percentage e.g. x is 50% probably bot

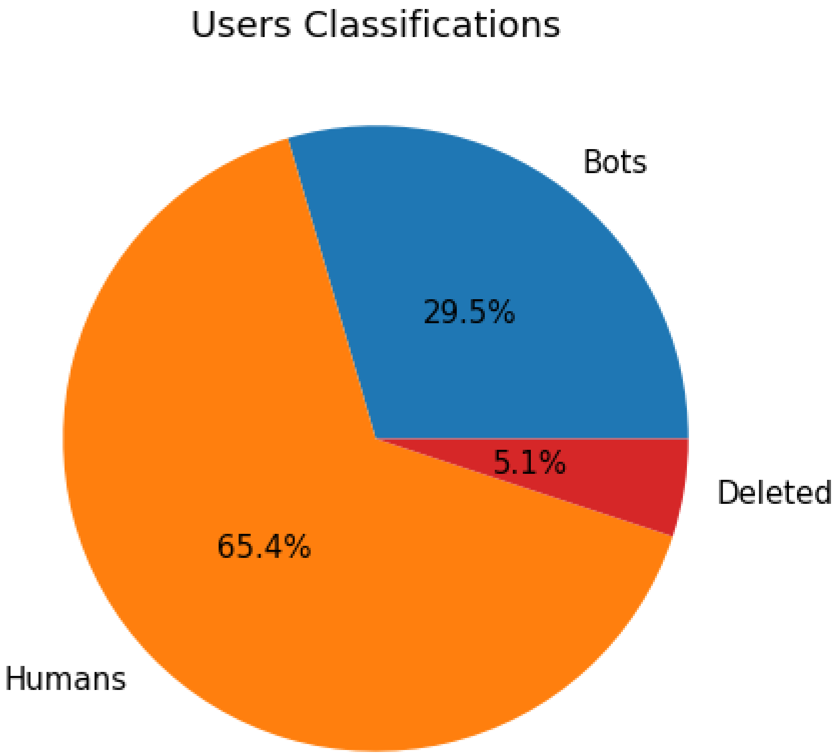
Data processing

Tweets_2020 data was more than 600k rows so it needed hard wrangling
First I set rows by created_at_1(i.e. most recent tweets) to got last tweet so when taking the difference with created_at (account creation date I got the age at time of dataset collection)
Then I used user.screen_name to remove duplicates users
I did some calculation as shown in notebook to got the missed feature
4 features used by machine was unavailable even with calculation so I used tweepy to got these features (favourites count,default profile,default profile image,geo enabled)
Not that the feauter gathered with tweepy are the most recent (i.e. now in June 2021) the rest were in 2020

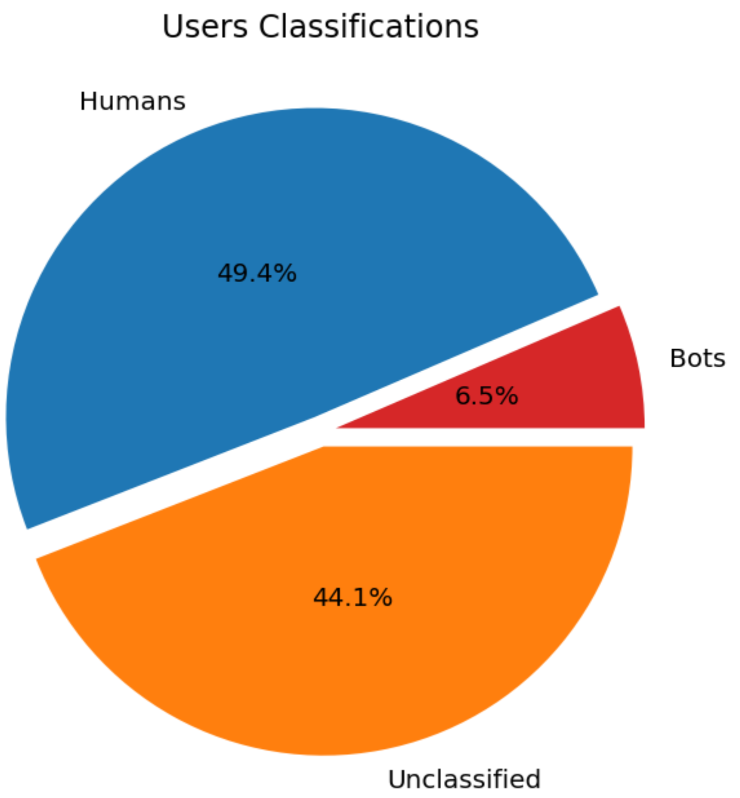
Results

Machine learning bots

1. About 33802 were classified as a bot from about 115k accounts
2. About 74861 were classified as human from about 115k accounts
3. The rest 5787 were deleted/suspended accounts when we searched for them using tweepy



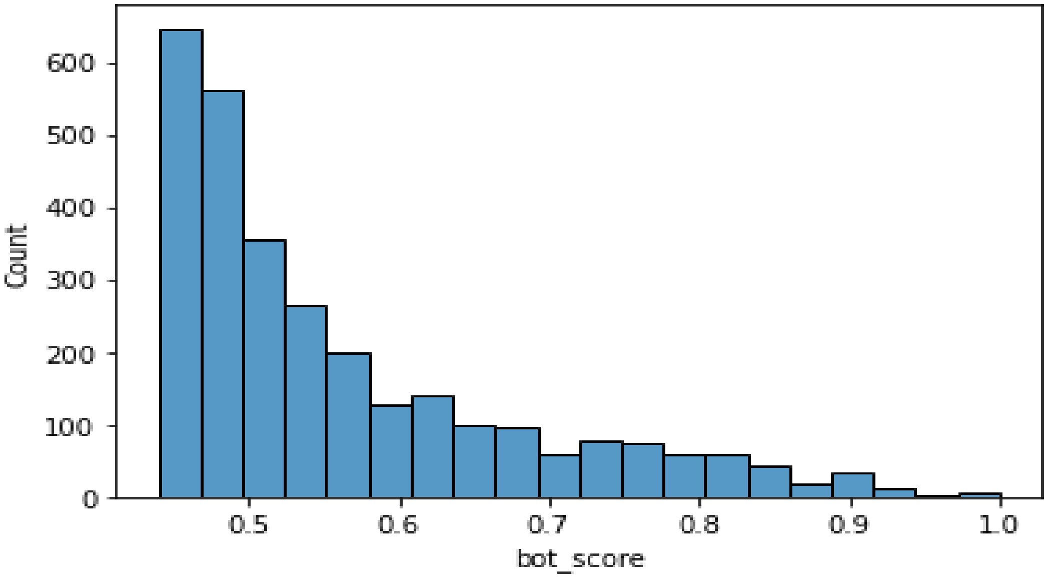
Botometer Analysis for 2021 Users



1. **1.** Total size of users = 48,000
2. **2.** Total size of bot users = 3175
3. **3.** Total size of Humans = 24225
4. **4.** Total size of unclassified users = 21593

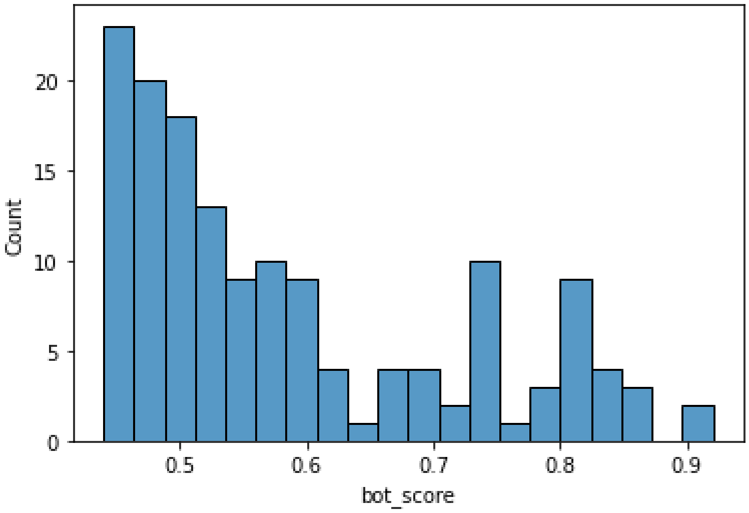
Combination of Botometer and ML Classifier Results

1. Botometer all account detected from 48k 2021 users = 27400
2. Botometer all bots equal 0% probability =616
3. Botometer all bots above 43% probability =2936
4. Botometer all bots above 50% probability= 1590
5. Botometer all bots above 70% probability= 413
6. Botometer all bots above 90% probability= 22



Botometer account

7. Bots with ML in ~ 115k accounts = 34053
8. Bots which are found also in Botomoter datasets= 853
9. Bots which are also in botomoter and above 43% = 149
10. Bots which are also in botomoter and above 50% = 94
11. Bots which are also in botomoter and above 70% = 34
12. Bots which are also in botomoter and above 90% = 2



Trends

1. Trend occur in 2nd Feb and from 6-13 Nov
2. I merged all trend in one dataframe with also bots dataframe
3. All trend are about 182169 tweet

4. About 15769 tweets created from bot (detected by ML model)
5. About 6270 tweets were created from account with less than 10 days age (account recently made in 24 Hour)
6. About 1349 account have made more than 10 tweet in the trends
7. About 93 account created in less than 10 days and tweeted more >= 10 tweet
8. About 492 tweets created from bot created less than 10 days
9. Jasmine Benedicto detected as a bot with ML model made about 14 tweets and created in about 1 day

Deleted Accounts

1. Number of deleted users from 2020 ~115k accounts = 5787
2. Deleted accounts which was also in botometer 2021 = 20 most probable bot was MERKURIUSME with probability of 53%